WHITE PAPER

NEH ODH DIGITAL HUMANITIES START-UP LEVEL I GRANT, 2015-2018

Notoriously Toxic: Understanding the Language and Cost of Hate and Harassment Online Games

Authors: Ben Miller (Project Director), Jennifer Olive, Cameron Kunzelman, Kelly Bergstrom, Wessel Stoop, Susan Benesch, Cindy Berger, Antal van den Bosch, Mia Consalvo, Kishonna Leah Gray, Todd Harper, Davin Pavlas, Nicholas Subtirelu

Abstract

An NEH ODH Digital Humanities Level 1 Start-up Grant from 2015-2018 supported an international workshop meeting and subsequent research aimed at developing approaches for studying and moderating online communication's frequently hateful, toxic and harassing speech. Operating under the title, Notoriously Toxic: Understanding the Language and Cost of Hate and Harassment in Online Games, this project gathered 10 scholars and practitioners representing disciplines in the humanities, social sciences, law, and media industries, and one industry partner, to focus on toxicity in the text-based chat systems of online games.

According to a 2016 report "47% of internet users have experienced online harassment or abuse,"[i] and "27% of all American internet users self-censor their online postings out of fear of online harassment."[ii] On a similar note, a survey by The Wikimedia Foundation showed that "38% of the 3,845 Wikimedia editors surveyed (estimated total over 130.000) had experienced some form of harassment, and over half of those contributors felt a decrease in their motivation to contribute in the future."[iii]

This multi-year collaboration focused on a two-day working group meeting of scholars from English, Linguistics, Law, Psychology, Education, Game Studies, and Justice Studies in consultation with industry experts from game development. It yielded this white-paper on, 1) best practices for studying and moderating toxicity, 2) conceptual and legal frameworks for addressing hate speech, dangerous speech, and toxic speech, 3) patterns of toxic language in online media, 4) next steps for building a reference corpus of toxicity types and a descriptive taxonomy, and 5) a humanistic perspective on consequences of toxicity and its moderation procedures. Practical anonymity in online communication has changed standards for interpersonal language; this project documents the most damaging of those changes. Among other project accomplishments were the securing of follow-on funding on this topic by Wessel Stoop, advisee of project participant Antal van den Bosch; the securing of a tenure-track position by project participant Kelly Bergstrom; and publication of a book chapter authored by a team including project participant Consalvo.

Contents

Project Participants

Miller, Ben. (PI) Senior Lecturer of Technical Writing and Digital Humanities, Affiliate Faculty in Quantitative Theory and Methods, Emory University.

Benesch, Susan. Faculty Associate at the Berkman Center for Internet and Society at Harvard, Adjunct Associate Professor at the School of International Service, and Director of the Dangerous Speech Project, American University.

Bergstrom, Kelly. Assistant Professor of Communications, University of Hawaii, Manoa.

van den Bosch, Antal. Professor of Communication and Information Studies and Centre for Language Studies, Radboud University Nijmegen.

Consalvo, Mia. Professor and Canada Research Chair in Game Studies and Design, Director of the mLab, Concordia University.

Gray, Kishonna Leah. Assistant Professor in Communication and Gender and Women's Studies, University of Illinois at Chicago.

Harper, Todd. Visiting Lecturer at the Simulation and Digital Entertainment Program, University of Baltimore.

Jenson, Jennifer. Professor of Pedagogy and Technology in the Faculty of Education, Director of the Institute for Research on Digital Learning, York University.

Kunzelman, Cameron. Doctoral Fellow, Communication, Georgia State University.

Lin, Jeffrey. Lead Designer of Social Systems, Riot Games.

Pavlas, Davin. Human Factors Psychologist & Game Researcher, Riot Games.

Subtirelu, Nicholas. Doctoral Fellow, Applied Linguistics, Georgia State University.

Notoriously Toxic

Overview

With the support of the NEH ODH and a Digital Humanities Start-Up Level 1 Grant, the "Notoriously Toxic" project under the direction of PI Ben Miller brought together 11 individuals representing 8 academic institutions and one game company to research methods for understanding and moderating on-line hate speech in video games. Over the course of the three year project, the team was able to accomplish the following goals:

- Develop a theoretical framework for understanding toxicity and online communities.
- Develop an ontology for data capture related to toxic speech events in the chat systems of online games.
- Collect, code, and analyze the community moderation guidelines for 105 games, game companies, and gaming platforms.
- Develop and implement a preliminary machine learning based classification tool to detect toxic speech in real time, and put forward a model for how to implement such a tool.

Background and Preliminary Findings

"Notoriously Toxic" presents a preliminary study of the language and impact of hate speech in the chat systems of online games. Developed by a group of researchers in game studies, computational linguistics, sociolinguistics, and law, and guided by an overall tripartite feedback model broadly corresponding to shielding potential victims from harm, educating those who casually engage in hate speech, and censuring those who persist in abusing their fellow players, the hope is that research-driven technical and social interventions might slowly shift online discourse norms away from casual, vicious, and potentially dangerous speech. Identification at scale of textual expressions of toxic behavior in online environments is a necessary, empirical preliminary aspect of this work to understand the prevalence and cost of online hate, as is qualitative cultural studies of the games and their player populations. A recent example of qualitative framing work in this area was the "Mapping Study on Projects Against Hate Speech Online" released in 2012 by the British Institute of Human Rights for the Council of Europe project, Young People Combating Hate Speech in Cyberspace (The British Institute of Human Rights Council, 2012). That report provides terminology and an environmental scan of processes aimed to limit hate speech online and offers suggestions as to new procedures. It, along with an examination of the reporting systems implemented across a host of online games, computational modeling of the language prevalent in these chat systems, and a study of work in the political sphere to defuse hate speech prior to its catalyzation of violence, serve as the foundation for this research.

Recent inquiries into the toxic elements of gaming cultures have primarily focused on communication outside of a game environment. For example, critical discourse analyses of player posts to online gaming forums found that heteronormative undertones of the World of Warcraft player community creates a culture of hostility toward LGBTQ communities (Pulos, 2013) and the same forum's adamant disavowal of feminism have made community conversations about gender roles and/or equality all but non-existent (Braithwaite, 2013). Similarly, Gray's (2012a; 2012b; 2012c) ethnography of Xbox Live demonstrates the constant barrage of gender and racially motivated harassment faced by women of color who opt to

communicate with teammates via voice chat. Finally, community leaders' adamant position of gender based harassment being a 'non-issue' is summarized by Salter and Blodgett (2012), whose case study of Penny Arcade's (a popular webcomic and organizers of PAX, a successful annual gaming convention) dismissal of its responsibility in perpetuating rape culture and SXSW Interactive's recent declaration that conversations about harassment in the games space can by definition not be civil (Sinders, 2015) is indicative of an industry that is highly resistant to change unless external pressure is applied. Taken together, this scholarship is evidence that toxicity exists across gaming culture writ large, and is not isolated to a particular game or specific player community.

Studying this phenomena at webscale and in the ephemeral environments of multilingual online chat systems is complex and requires a multidisciplinary approach bridging core strengths in the humanities, such as cultural criticism, with strengths in social psychology, the data sciences, and linguistics. Studying the socially destructive behavior as manifested in online gaming platforms encourages innovative approaches to this problem. One corpus examined as part of this research is comprised of the chat logs produced by the player base in Riot Games' League of Legends (League). As of January 2014, League had ~27 million unique players every day each playing no less than 20 minutes and a peak concurrency of 7.5 million people who collectively have logged billions of hours of total play time for the game since 2009 (Sherr, 2014). Given that the game is a global phenomenon, the chat logs contain harassment in virtually every language.

Based on the UN framework provided in the International Covenant on Civil and Political Rights (1976), Susan Benesch generally defines hate speech as ". . . an expression that denigrates or stigmatizes a person or people based on their membership of a group that is usually but not always immutable, such as an ethnic or religious group. . . . Speech may express or foment hatred on the basis of any defining feature of a minority or indigenous people, such as ethnicity or religion – and can also denigrate people for another 'failing,' such as their gender or even their location, as in the case of migrants" (Benesch, 2014, pp. 20). This broad but inclusive definition is further elaborated upon by Nazila Ghanea in reference to the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) in the establishment of a spectrum from least to greatest: discriminatory speech, hate speech, incitement to hatred, incitement to terrorism, and incitement to genocide (Ghanea, 2013, pp. 940-1). The characteristics of these definitions reflect the significant impact that hate speech acts have on the establishment and enforcement of personal and communal identity and the need to identify such acts in order to preserve those identities. In his discussion on Carey's ritual model of communication as applied to cases of hate speech, Clay Calvert explains how hate speech initializes and perpetuates the subordination of one group over another (Calvert, 1997). Calvert notes that hate speech acts, specifically focusing on the repeated utilization of racial epithets, construct reality in the speaker, audience, and target members of the discourse through the creation and maintenance of mental schemata similar to the functions of other speech acts: "In particular, racist speech helps to define who minorities are and how others think about minorities, facilitating their unequal treatment" (Calvert, 1997, pp. 12). This construction is harmful in its immediacy to the target as well as in the long-term situation as it perpetuates unequal power structures based on criteria of identity (Calvert, 1997, pp. 15-16). The construction of reality based on hate speech acts is also relevant to a discussion of online environments as the textual communication serves as a large social aspect of both on- and offline environments.

Toxicity consists of verbal expressions and behaviors that serve to destabilize groups. It is unclear whether toxic behavior is directed to elicit particular responses, and hence systemic, or reactionary, emotional, and hence, situational. Frequently, the term 'troll,' or 'trolling,' is used synonymously with toxicity. Regardless, these behaviors are necessary to address because they are a key factor in the outright hostility of online gaming environments to those perceived as other. This destabilization maps on to offline models of gender, race, class, ethnic, national, linguistic, and abelist-based hate speech. A fertile ground for this analysis is in the virtual worlds of online gaming. For example, Kou and Nardi (2013), in their research on League of Legends, found that antisocial behavior destabilizes online communities but is addressable by social code and regulatory systems.

The targeted examples in online gaming environments show a concentrated sample of toxic behavior that is pervasive in every online environment. Studies have documented antisocial behavior and toxic speech in most digital platforms (O'Sullivan & Flanagin, 2003). Whether considering flaming on 1980s and 90s USENET forums, social media fueled outrage on contemporary politics, or in-game, text-based conversation in multiplayer games, toxicity can be motivated by emotional, intellectual, political, or other causes and as such correlates strongly with the modes and consequences of offline hate speech. Understanding online toxic behavior in ways that allow for moderation of its causes and effects first requires an understanding of how to study the concrete manifestation of this behavior—the text produced by users of a system. The challenges faced by this research are many: the writing styles are heavily infused with jargon, the orthography is non-standard, the chat stream only represents one channel of communication, and the communities are fluid.

In response to these challenges, an approach grounded in machine learning and NLP was tested. Using a small subset of the available data, a classifier based upon developed to separate players toxic from non-toxic players yielded a precision of 0.77, a recall of 0.79 and an F-score of 0.78. These results are encouraging, and along with training on a larger data set, secondary factors such as player avatar gender, length of match, and others were also preliminarily tested and found to have small influencing effects. These results suggest that there are concrete, detectable semantic and syntactic patterns in the harassment levied at players in these games. Connecting these findings to mechanisms for shielding, reforming, and censuring players, and to frameworks for understanding the social and psychological costs of being effectively locked in a room with one or more individuals determined to verbally abuse a peer is the more complex task of the cultural and ethnographic studies of digital communities.

A Theoretical Framework for Understanding Toxicity at Different Scales

Attempts to document sociality in online gaming communities have taken multiple forms. Here, we outline a brief overview of the literature relevant to the study of toxicity and anti-social interactions roughly divided into three levels of scope: micro (individuals), meso (groups), and macro (entire servers/communities).

Micro-level

Typically, micro-level investigations in game studies uses an individual player as the unit of analysis. Studies of individual players often rely on interviews (Valkyrie, 2011; Whitty, Young, & Goodings, 2011) and surveys (Decker & Gay, 2011; Li, Liau, Gentile, Khoo, & Cheong, 2012; Liu & Peng, 2009; Suznjevic & Matijasevic, 2010) with the goal of learning more about how MMOGs fit into an individual player's life. This research is

predominantly grounded in psychology and can be grouped into two large categories: studies of the "escapist" qualities of MMOGs, and studies of what sorts of personality types are attracted to this genre of games.

For studies related to the investigation of toxicity at the micro-level, we turn to investigations of personality types or what has become known as "player type". This is a body of research that seeks to determine who plays games (most often MMOGs) and for what reasons. Frequently referenced by players and academics alike, Bartle's "Hearts, Clubs, Diamonds, Spades: Players who suit MUDs" (1996) is used to explain what motivates an individual to play a MMOG. Bartle's taxonomy, created from his experience as a game designer, describes four primary motivations for playing MUDs: achieving, exploring, socialising, and killing. This typology of player motivations – especially the "killer" or griefer – has been applied descriptively to multiple MMOGs and virtual worlds including World of Warcraft (Ang & Zaphiris, 2010; Paul, 2010), City of Heroes/Villains (Meyers, 2007), Second Life (Chesney, Coyne, Logan, & Madden, 2009), and Ragnarok Online (Lin & Sun, 2005). These investigations seek to uncover why some players prefer to disrupt the games of others rather than participating in pro-social or achievement-oriented behavours.

Bartle's taxonomy remains popular because it provides a simple to use model offering plausible explanations for the different types of behaviors that can be observed among MMOG players. This model, however, was not subjected to empirical testing until Yee (2006) conducted a large online survey based on Bartle's descriptions. Yee's survey resulted in a reworking of Bartle's model to expand on the number of possible motivations to allow for more nuances and variation when discussing what players might get out of MMOGs (see Figure 1). Bartle's typology is still visible, but while his model only allowed for a player to be slotted into a single category at the exclusion of all other categories, Yee's updated version anticipates that players will be slotted into multiple categories according to their survey responses. While Yee's model represents a step forward as it allows for greater flexibility, it is important to note that both his and Bartle's earlier model were conceptualized with fantasy-themed role-playing games in mind and may not be as easily applicable to the other genres of games discussed throughout this report.

TABLE 1. SUBCOMPONENTS REVEALED BY THE FACTOR ANALYSIS GROUPED BY THE MAIN COMPONENT THEY FALL UNDER

Achievement	*Social*	*Immersion*
Advancement Progress, Power, Accumulation, Status	**Socializing** Casual Chat, Helping Others, Making Friends	**Discovery** Exploration, Lore, Finding Hidden Things
Mechanics Numbers, Optimization, Templating, Analysis	**Relationship** Personal, Self-Disclosure, Find and Give Support	**Role-Playing** Story Line, Character History, Roles, Fantasy
Competition Challenging Others, Provocation, Domination	**Teamwork** Collaboration, Groups, Group Achievements	**Customization** Appearances, Accessories, Style, Color Schemes
		Escapism Relax, Escape from Real Life, Avoid Real-Life Problems

FIGURE 1: YEE'S UPDATED PLAYER TYPES. YEE'S MOTIVATIONAL FRAMEWORK CONTAINS MANY OF THE APPEALING ELEMENTS OF MMOG PLAY, GROUPED UNDER THREE "UMBRELLA TERMS." CHART IS REPRODUCED FROM YEE (2006, P. 774).

With Yee's model now widely adopted alongside Bartle's original taxonomy, researchers have begun to cross-reference survey responses with in-game observational data to further refine and elaborate these models. For example, Suznjevic and Matijasevic (2010) were interested in combining the *why* players play (i.e. their motivation) with what players *do* when they play (i.e. their preferred in-game activity). Using survey and observational data, Suznjevic and Matijasevic find that player types are linked to specific in-game activities. The strongest correlation between player type and particular in-game activities was seen among players that are motivated by advancement-oriented goals are more likely to participate in end-game raiding (p. 18). Yee has also been involved in such cross-referential analyses, combining survey work with data-logs of *EverQuest II* player activities (Williams et al., 2008), but to date his motivational framework remains unchanged from the version included in Figure 1.

Testing player types by means of Yee's framework has been recently been critiqued by Li et al. (2012) for being too detailed and requiring too many survey questions to properly assess a participant's motivation(s) for play. Citing a fear of participant fatigue and a low survey response rate, Li et al. have attempted to distil Yee's motivational framework and its associated assessment (39 survey questions) into a truncated survey (12 questions). Their stated intent for shortening the survey is to create an easy to use motivational framework to better identify player types who are more likely to exhibit pathological or "at-risk" behaviors. Li et al. argue that players who seek immersion from their MMOG gaming experiences are the type of player most likely to develop problematic gaming behaviors, but in this case it is a warning that online interactions can become addictive, rather than leading to participation in toxic behaviors such as harassment or griefing other players.

Meso-level

Moving on from the study of individual players, game scholars have also investigated groups of players. This has ranged from small, temporary groups (Eklund & Johansson, 2010, 2013; Nardi & Harris, 2006; Nardi, Ly, & Harris, 2007) to studies of more permanent social structures like guilds. There is a lack of literature about toxic behavior in guilds, likely because these social structures act as a form Foucauldian surveillance in action (Chen, 2008, 2012; Silverman & Simon, 2009). Therefore, this review will focus on qualitative observations of temporary groups where data is typically collected through ethnographic methods. Participant observation is a key method for data collection at the meso/group level.

Research at the meso-level provides evidence that sociality in games can be heavily influenced by developer decisions in regards to how players can come together in temporary groups. Recently, Eklund and Johansson (Stetina, Kothgassner, Lehenbauer, & Kryspin-Exner, 2011, p. 477) traced the implementation of a new feature (the "random dungeon finder") within *World of Warcraft* that allows for the easy creation of 5-person Pick Up Groups (PUGs), but has had the unintended consequence of reducing overall social interaction among players. Drawing comparisons between Goffman's (1990) descriptions of encounters and their own in-game observations, Eklund and Johansson observed the frequency and content of conversation between PUG members with the goal being to determine the norms and expectations for social interaction in this type of group formation. The random dungeon finder is a way of queuing yourself for a particular role in a dungeon group (tanking, damage, or healing) and the game software then finds complementary players for you to be grouped with. This tool is quick and efficient, but comes with a trade-off: most players grouped

together by this tool will be from different servers and will never have the opportunity to group together again. Eklund and Johansson participated in 24 of these random groups, recording conversations that took place in the dungeon and observing the interactions between players. What they found was unsurprising – players from different servers did not participate in small talk or conversation. What *was* said was predominantly instrumental, using the minimal amount of text required to exchange information to efficiently keep the group moving towards their final goal. It was rare to find players who knew each other prior to entering the dungeon and mostly then that conversation could be characterized as social rather than instrumental. These prior connections were infrequent and most players were assumed by the authors to be aware that as soon as the dungeon was completed, they would never see other players again. Eklund and Johannson argue there is little incentive to create a social bond or perform supportive rituals when participating in random groups. Instead, other players become tools to complete a specific task (in this case, a dungeon) rather than viewing group members as potential friends.

That a game's affordances directly influence the tone of social interactions among strangers is most apparent when Eklund and Johannson's findings are contrasted with work done about temporary groups prior to the implementation of the random dungeon finder. In Bardzell et al.'s (2008) study of 5-person PUGs in *World of Warcraft*, researchers sat beside an informant and observed their participation in a dungeon group, and then interviewed the player after the group disbanded about what, in their opinion, made for an enjoyable group experience (p. 358). While the in-game group formation tool described by Ekuland and Johannson allow players to place themselves in a queue and complete other tasks while the game software sought out other party members on their behalf, Bardzell et al.'s informants had to go through a much longer, involved process before ever setting foot inside a dungeon. They describe the multiple steps in this time consuming process:

> • Recruitment: Players assemble a team of five; this stage is often the most difficult and frustrating (indeed, two of our six teams failed to go beyond this phase, though they each tried for nearly an hour).
>
> • Pre-instance planning session: Players agree on roles (healing, melee, support, etc.), behavioral protocols, and looting policies.
>
> • Episodic skirmishes: The group faces small groups of enemies in brief battles, followed by a short rest.
>
> • Boss battle: A couple of times per instance, the group will encounter boss characters.
>
> • Loot distribution: After the boss is defeated, loot is distributed, often following negotiations, mentoring, and other discussion.
>
> • Dissolution: The instance complete and loot obtained and distributed, the group disbands, begins a new instance, or continues elsewhere in-world. (Bardzell et al., 2008, p. 358)

Group formation is a rather arduous process, and their informants expressed a preference for playing with those previously known to them, as strangers tended to be less forgiving of mistakes (p. 359). At the time the data for this paper would have been collected, players were only able to group with others

located on the same servers as themselves. With only a limited pool of players to draw potential group members from, it follows that players would be concerned about being labelled as "lousy", and feared being blacklisted from future groups. Bardzell et al.'s look at temporary group formation describes a time before the random dungeon finder, a time where personal relationships were key to successful dungeon encounters, which is a sentiment lacking in the encounters described by Ekuland and Johannson. For our purposes here, it is interesting to note that players were more concerned with being labeled as having game-playing skills considered to be bad or "lousy" rather than fearing social shunning for being a toxic or a negative social force in the group.

Similar to Bardzell et al., Nardi and Harris' (2006) early look at *World of Warcraft* describes the utility of temporary groups outside of dungeons as a means for players to complete a difficult quest, but also describe how the features of the game software allow players to easily extend a successful collaboration well beyond the original impetus for banding together:

If party members enjoy playing together they may share their other quests. Quests are normally obtained from a computer character but quest-sharing allows players to continue to play together after the initial quest that brought them together is complete. If they like one another, players add each other to their friends list for future play. (Nardi & Harris, 2006, p. 152)

These short-duration temporary groups, Nardi and Harris argue, meet the characteristics of "knots" -- groups of strangers that come together to collaborate on a specific task (p. 154) -- and successful completion of this task will likely lead to further collaborations, or even the beginning of a friendship. More dramatic is the impetus for collaboration through "crisis scenarios" in *EverQuest*, as described by Yee (2008). These unexpected encounters with a hostile monster were designed to catch players off guard. Additionally, these monsters were powerful enough to necessitate all players in the immediate area to work together to defeat them. These unexpected attacks, combined with the harsh penalty for death in *EverQuest* forced players to work together. Unlike the random dungeon finder described by Ekuland and Johannson where players can click a button to leave the group at any time and immediately be transported to safety, the crisis scenarios described by Yee forced players into interactions that had no easy exit. Successful teamwork to defeat the crisis scenario could lead to a sense of camaraderie, and in turn feasibly result in a new friendship.

Temporary groups in *World of Warcraft* or *EverQuest* were previously shown as providing the opportunity to meet strangers and possibly create new friendships (Nardi & Harris, 2006; Yee, 2008). The in-game mechanics for creating these groups has since been modified, and according to Eklund and Johannson, no longer encourage the development of ongoing social bonds. The contrast between these examples are helpful for showing the ways in which the affordances of a particular MMOG can influence the types of social relationships possible within its gameworld. In these specific examples pertaining to MMOGs it is interesting to note that temporary groups (i.e. being in the same space for a limited amount of time to achieve a shared goal) does not necessarily lead to examples of toxic interactions. Instead, as Blizzard changes its game to allow for temporary grouping with strangers (rather than players forming groups on their own) it is an overall lack of social interaction (pro or anti-social) that tends to be observed.

One noticeable exception comes from EVE Online. EVE is a game that is positioned by developers as a "sandbox style" game and celebrated as an environment that fosters emergent play. "Games" within the game, such as the now annual event "Hulkageddon" started with a handful of players and now grown to the point that it impacts the gameplay for all *EVE* players (Bergstrom, 2011, 2016). Similarly, a group of players banded together under the corporation GoonSwarm, changing PVP so there is no longer a combat advantage for those who have been playing *EVE* longer than others. While GoonSwarm has been lauded by some for "levelling the playing field", others have critiqued their in-game activities for being nothing more than bullying and reducing the openness of the sandbox to a very limited type of play that is not attractive to other groups of players (including women). These are reminders that play, especially in a MMOG, is constantly shifting and changing.

Macro-level

Finally, beyond individual players and the studies of groups/communities, is the analysis of entire game servers. While micro and meso studies are comparable to studies of casual pick up games at the park and more formally structured sports leagues, these server-wide samples are more akin to the study of a small town. However, in this small town every interaction between townsfolk and the town's infrastructure is recorded and accessible through a central database. The motivation for this sort of research is to model large group behavior with the goal of identifying community-wide patterns and trends, rather than focusing in on the activities and behaviours of individual players.

One benefit of quantitative modelling of community-wide patterns is the ability to conduct research that may otherwise present "more than minimal risk" to those being studied. As previously mentioned, games like *World of Warcraft* have been used to model the spread of infectious diseases throughout a large population of actual people but without the risk of infection or death (Balicer, 2007; Boman & Johansson, 2007; Lofgren & Fefferman, 2007). The same game has also been fruitful for network analysis, such as studies using extensive logs of in-game text chat to learn more about who communicates with whom and for what goals (Ang, Zaphiris, & Wilson, 2010; Ratan, Chung, Shen, Williams, & Poole, 2010; Williams et al., 2006). Norms and expectations about what sort of armour and weapons each class is expected to equip was collected by scraping data from the *World of Warcraft* Armory, a database where information about the activities all active avatars is updated daily (Lewis & Wardrip-Fruin, 2010). Armory data has also been used by Yee et al. (2011) to explore norms surrounding avatar gender and specific roles within the game, specifically if stereotypes about male and female roles in MMOGs actually play out in the *World of Warcraft* player population. The datasets of these described papers are frequently large enough to be a generalizable sample of the player population, or in some cases, include data from the entire player population of the MMOG under investigation (Ducheneaut, Yee, Nickell, & Moore, 2006; Shen, 2010).

While MMOGs may appear as being an attractive petri dish for unobtrusively collecting data about large groups of MMOG players, the ability to relate these observations to society writ large is dependent on the principle of "mapping" being true. With its roots in economic and educational research, mapping seeks to determine if the behaviours observed within online worlds are similar, if not the same, as those behaviours that can be observed in analog situations in the offline, "flesh and blood" world (Williams, 2010, p. 452). Still in its early stages, the mapping principle has yet to be experimentally validated, so researchers

must still cross-reference data collected from MMOGs with data that can be used to verify a player's offline characteristics, such as demographic data collected via survey, or asking participants to share their Facebook profile. Proponents of mapping argue that this sort of work has advantages over the smaller scale studies discussed earlier in this review, namely that when studying an entire MMOG population there is no risk of sampling bias as all players are included in the sample, and with data being collected at the server level, players are unaware that they are being studied, and therefore less likely to change their behaviour as they would if they know they are being observed by a researcher (Williams, 2010, pp. 464–456). Concerns about lack of consent by members contained in these "big data" sets have been raised by de Castell et al. (2014; 2012) in regards to MMOGs and Zimmerman (2010) in regards to Facebook, but the ethics and legality of such research currently remain largely unaddressed by proponents of this research methodology.

Public Reception

One note worth mentioning is that this project almost instantly generated public attention. Within days of the NEH's announcement of the award, and among the first articles to be published in response was, "FEDS GIVE $29,000 GRANT TO CRAZY RACIST WHO CLAIMED INTERNET TECHNOLOGY IS "INHERENTLY WHITE AND MASCULINE." Appearing in *FrontPage*, an online magazine edited by David Horowitz and published by the David Horowitz Freedom Center, the article misidentified the project personnel, choosing to focus on an attack on one of the project's participants who happened to be a person of color, rather than other members of the project team who were not. The Southern Poverty Law Center noted that Horowitz, "has since the late 1980s become a driving force of the anti-Muslim, anti-immigrant and anti-black movements."[iv] They go on to note that, "FrontPage, which is still in operation, has become a platform for publishing a plethora of far-right and anti-Muslim writers and commentators," and that FrontPage, in 2002, was used to reprint material first published in *American Renaissance*, a white nationalist publication. In our conversations at the working group meeting, it was noted that each of the game studies researchers at the table had been direct targets of on-line harassment ranging from doxing (the publication of personal details online) to direct threats via email and other means. Conversely, none of the human rights researchers, individuals who have worked in contexts ranging from the prosecution of war criminals to election violence, had been the recipients of such direct threats. The study was also highlighted by a sitting congressperson, Steve Russell (R-OK 5th District) in his "Waste Watch" publication as an example of federal waste. "Notoriously Toxic" and its $29,403 award appeared as the second targeted item in his list, between a $432 million streetcar project and a $14.7 million military warehouse. In Russell's write-up of the project, he agreed with the necessity for this work and its goals: "Online threats, incitement, stalking, and harassment are certainly important issues that deserve serious study. The researchers would be best advised to focus on these aspects of their project," and quoted an author from a gaming news website that was closely affiliated with #GamerGate, a program of coordinated harassment of women in gaming; "Yes, there are horrible individuals who say horrible things in every online community, but I've not yet seen evidence that gamers are worse than any other group."[v] **This confluence of negative, racist attention struck the project team as evidence of the toxicity that requires projects like this one.** Identifying and studying hate speech strikes a nerve.

Follow-on Funding

Since the conclusion of the working group meeting, multiple proposals for follow-on funding were submitted to the NEH, Facebook's Protect and Care Division's "IT and Society" RFP, and the NWO-funded Language in

Interaction Consortium. Those proposals have looked to expand the global reach of this project and automate some of its findings. One project, "Toxicity and On-Line Radicalization in News and Games from the Middle East and North Africa Region: A Multi-lingual Analysis of Temporal and Conceptual Metaphor Patterns" (TOLiR-MENA), looked to include toxicity in comments to newspaper articles, thereby expanding the domains under address, do so in both English and Arabic, thereby expanding the languages being studied, and study platforms localized to North America, Europe, and the Middle East and North African regions, thereby globalizing this work. Another project, "HARE: A tool to automatically recognize online harassment in progress," aims to automate theoretically grounded methods for identifying toxicity in real time.

Machine Learning and Hate Speech

Collective conversations at the workshop and thereafter put forward, starting in 2014, that real-time machine-learning-based applications can detect digital harassment as it happens. Software offering new functions can then be plugged into new or existing software such as games or online chat environments to give automated judgments on the 'toxicity level' of the various participants in online conversations. The potential impact of this type of solution is that owners of conversation platforms obtain a new and practical tool to detect digital harassment in process automatically. Based on this information, they can (if desired automatically) take action to keep the conversations respectful, professional, insightful, useful and/or pleasant to all other participants - whatever the goal may be. The most effective counter strategy will vary from platform to platform, but could for example be notifying or muting the harassers, or signaling to non-toxic participants that they can flag toxic participants if they feel harassed.

Although such a tool could be set up in such a way that it can be trained on any domain/language combination and therefore be language-independent and domain-independent, our focus has been on toxicity in videogames. Using data from the League of Legends game, our team developed classifiers that identified some types of toxic speech in real-time. More complex rhetorical movements, such as sarcasm, escaped detection, but more common "name-calling," or "direct threat" style toxic speech was readily apparent. Further work is required before that research can be published as it was beyond the scope of the proposal for this project. Two outcomes of this project aimed at facilitating that research are an ontology for data related to toxic speech in game chat-systems, and a survey of community moderation policies across the game industry.

Toxic Speech Data Ontology

Developed principally by project key personnel Antal van den Bosch and PI Miller in consultation with the entire team, the following ontology describes, broadly, the kids of data necessary to assess the discursive and rhetorical situation of online hate speech. It incorporates details about the messages themselves, the systems in which they exist, and the entities responsible for their construction. With this type of data, a project such as "Notoriously Toxic" would be able to identify more incidents of toxicity than the aforementioned "name-calling" and "direct threat" types, do so more accurately, and perhaps do so in more complex rhetorical framings such as via sarcasm and irony.

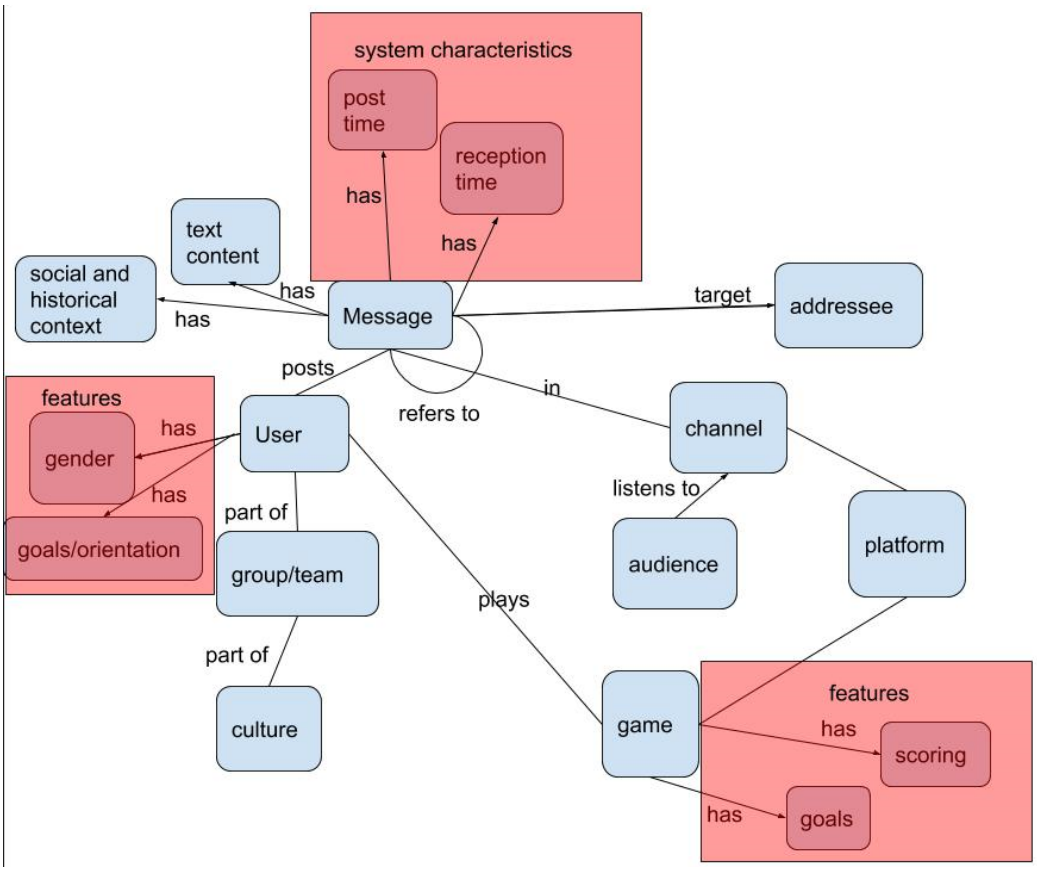

FIGURE 2: PRELIMINARY ONTOLOGY FOR TOXIC SPEECH DATA

Community Moderation Policies

This project element was led by project key personnel Kelly Bergstrom, and was produced via the efforts of Jennifer Olive, Cameron Kunzleman, project PI Miller, and Cindy Berger. Our goal in undertaking this survey was to understand, based on the frameworks of the community, what was considered toxic. In addition, we were interested in the "scales" at which these determinations were made, and what consequences the community considered possible and legitimate. After collecting the community moderation guidelines from 105 distinct sources covering the gamut of contemporary games (e.g. World of Warcraft), game platforms (e.g. Xbox Live), and game companies (e.g. EIDOS), we qualitatively coded the corpus using an iterative coding methodology across 5 coders and three passes. The first pass was to develop individual coding schemas which were then compared and normalized. After that step, a second coding pass was conducted using the new coding schema. Following comparison and discussion, the final coding schema was developed and then applied to the corpus as a whole. The two-level schema is presented in Table 2 below. The results of the study will be released at a later time.

TABLE 1: QUALITATIVE CODING SCHEMA FOR COMMUNITY MODERATION POLICIES

Coding Category (L1, L2)	Description
Behavior	Identified toxic behaviors including but not limited to speech and actions
Decisions	Outcomes of procedures for behaviors determined to be toxic (i.e. muting, re-naming, temporary ban, permanent ban, etc.)
Definite Legal Action	
Mute	
Other	
Permanent Ban	
Potential Legal Action	
Re-name	
Temporary Ban	
Toxicity Counter-behavior	i.e. ignore toxic player, present a positive attitude, etc.
Domains	Place in which the toxic behavior takes place
Account	
Forum	External forum to the game
Game	Behaviours occurring in-game.
Other	
Outside Platform	
Website	
Jurisdiction	Where is the decision operative?
In-Distribution Platform	i.e. Steam
In-Game	
In-Game Related Communication Channels	i.e. Twitch.tv or Reddit forums
In-Publisher Platform	i.e. Battle.net
Other	
Out-of-Game or Publisher Action	
Procedures	Ways/mechanisms of determining toxic behaviors (i.e. tribunal, auto-detection, etc.)
Auto-detection	

Coding Category (L1, L2)	Description
Carte Blanche or Developer or Publisher Exemption	
Moderation or Staff Discretion	
Other	
Reporting	
Tribunal	

The data resulting from this coding operation is summarized below in figures 3-7. Figure 3 shows the relative proportion of each of the five code families and the most commonly occurring sub-codes. For example, under the Decision family of codes, the most common reference within the documentation describing moderation procedures across 105 games, franchises, systems, and publishers, was to "permanent bans." "Muting," a punishment that permanently or temporarily restricts a player's access to chat, was the smallest contributor to that category.

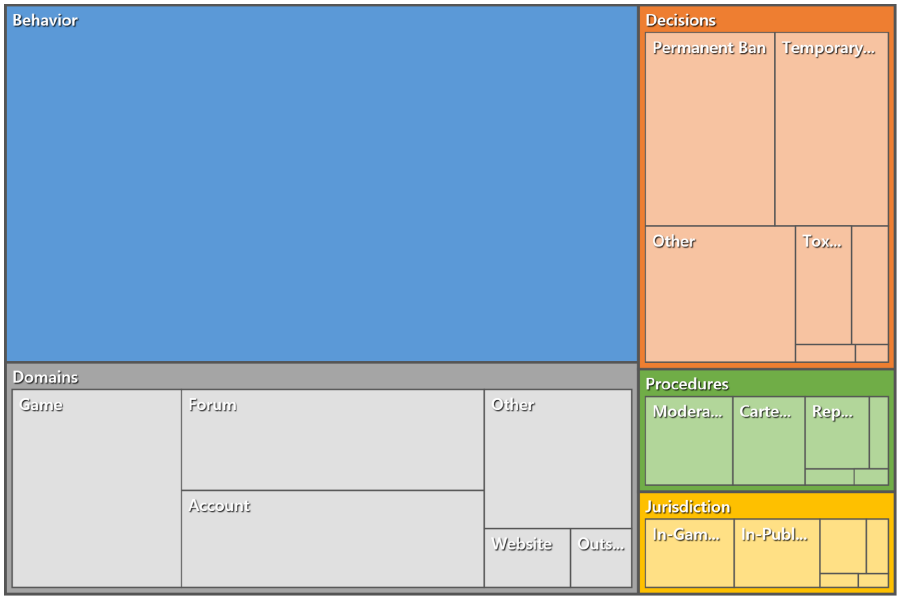

FIGURE 3: RELATIVE PROPORTION OF CODES AND SUB-CODES FOR ALL 105 CODED DOCUMENTS

Figures 4, 5, 6, and 7, respectively, indicate the relative percentage of a particular document in which a given family of codes is discussed. The percentages are not mutually exclusive nor do they need to add up to 100%. This is because a given moderation standards document may discuss procedures in respect to a given punishment decision within a particular domain. In which case, each of those families would be operative within that section of a document. Figure 4 indicates that the "Nintendo Code of Conduct" addresses what constitutes an actionable violation of behavioral standards in 89% of its content, versus, for example, Microsoft's Code of Conduct that does so in 25% of its content. Figure 4 highlights that a greater percentage of the documentation about League of Legends, as contained in their "Summoner's Code"

document, as opposed to Blizzard's documentation about their Battle.net service, deals with "Decisions," aka punishments.

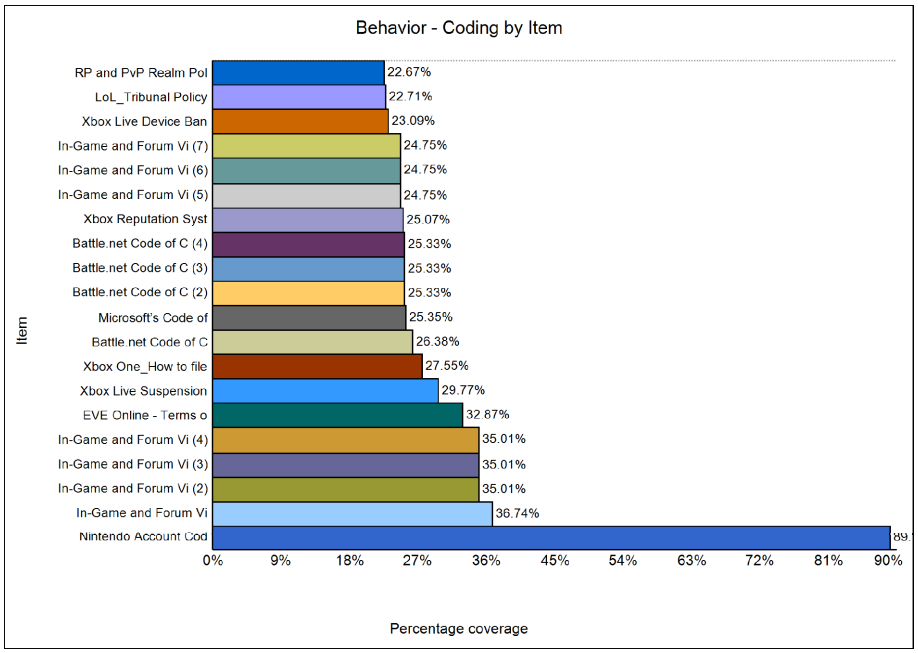

FIGURE 4: PERCENTAGE OF DOCUMENT ADDRESSING ACTIONABLE BEHAVIORS

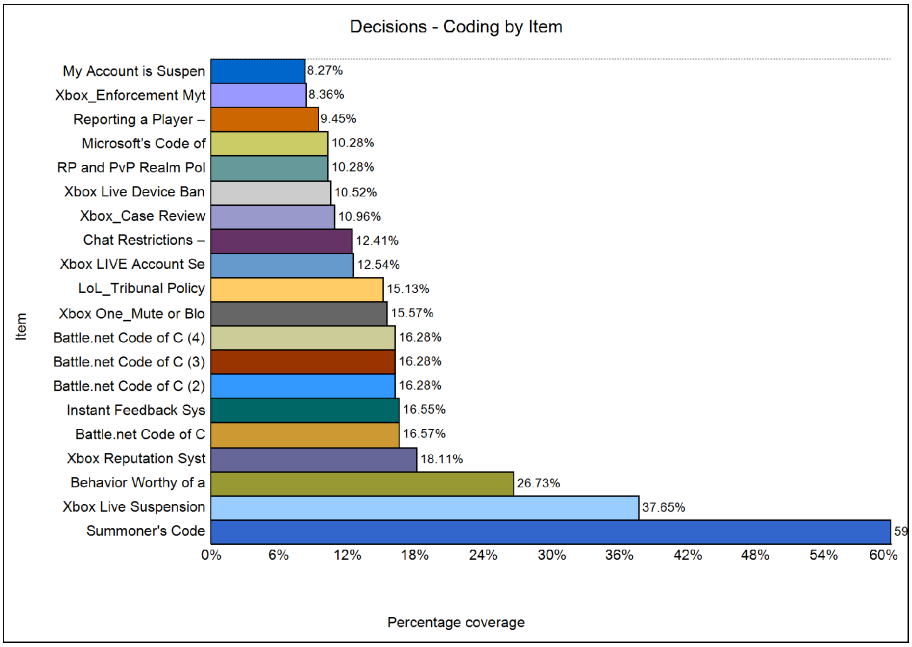

FIGURE 5: PROPORTION OF DOCUMENT DESCRIBING "DECISIONS," I.E. TYPES OF PUNISHMENTS

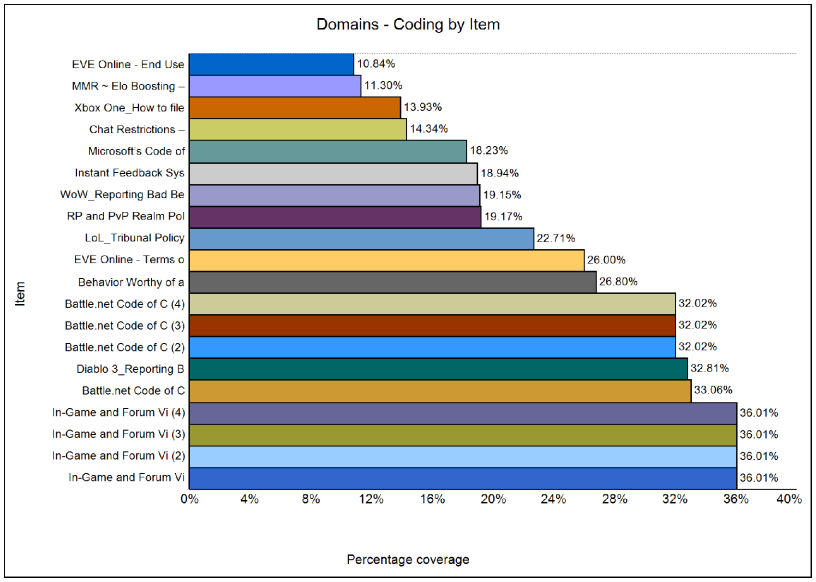

FIGURE 6: PROPORTION OF DOCUMENT PERTAINING TO DEFINITIONS OF DOMAINS WHERE MODERATION CAN TAKE PLACE

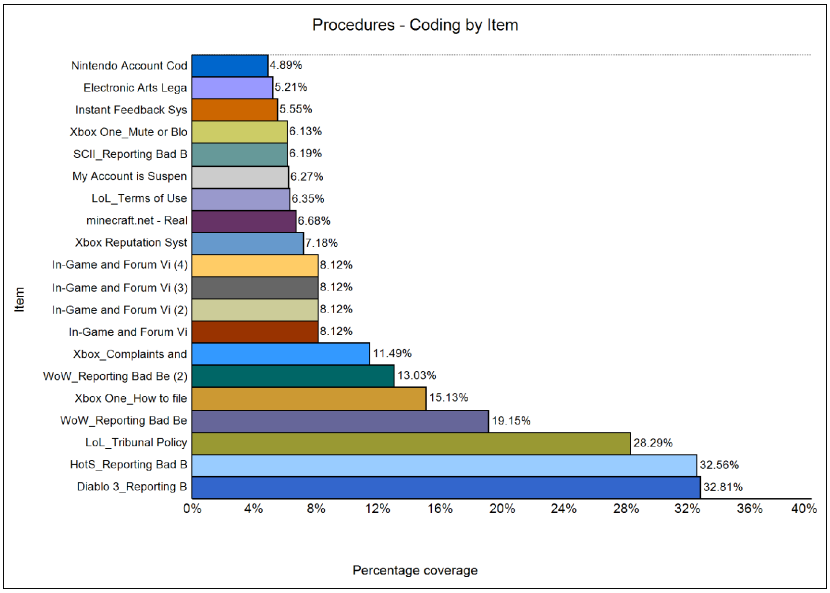

FIGURE 7: PROPORTION OF DOCUMENT CONTAINING CODES RELATED TO DECISION MAKING PROCEDURES

Conclusion

When chat systems are integrated in games the toxicity problem tends to be exacerbated. Although there is no definitive research as to why, conjecture suggests that it may be because players are often required or compelled to stay in the game despite harassment. Common social strategies for dealing with bullying, such as walking away, are not available in games. In addition, games are common spaces that build communities of practice that cross demographic, linguistic, and geographic boundaries. In particular, we will focus on the 'Multiplayer Online Battle Arena' (MOBA) video game genre, a relatively new genre that is

enjoying great successes: in the list of most popular online videogames, we see MOBAs League of Legends, DOTA2 and Smite at positions 1, 3 and 6, respectively. Other options would include “Battle Royale” style games like Fortnite. Each of these games report multiple millions of players per month, with League of Legends topping the list with an estimated 100 million. MOBA games share a number of game mechanics that incentivize strategic team play, but they also share severe problems with so-called “toxic players.” Toxic players deliberately insult or harass teammates and/or opponents, thereby ruining the game for all other players in the match.

Since this project was first proposed in 2012, interest in this area has grown. That interest is a by-product of the increased stakes of online toxicity. As more professional, personal, and cultural life has shifted into semi-public online forums, the possibility has grown for disruption in those forums to disrupt the essential activity they support. Toxicity in games has been linked to tragic, real world violence. It has been linked to the suppression of ideas. Unfortunately, toxicity is a feature of online communication, and there are no solutions that do not require work.

We are grateful to the NEH ODH for supporting this contribution to the understanding of the problem of online hate speech, and the development of solutions addressing attempts to fracture these nascent, vital online communities. The experience of working among this team of 11 energetic, dedicated scholars and game developers was enlightening and invigorating. It is our hope that this project represents one step forward in a broader effort to understand and address the challenges posed by pervasive online hate speech, both as it appears in the games we play, and in the other forums upon which our communities and identities increasingly rely.

Works Cited

Ang, C. S., & Zaphiris, P. (2010). Social Roles of Players in MMORPG guilds. *Information, Communication & Society*, *13*(4), 592–614. http://doi.org/10.1080/13691180903266952

Ang, C. S., Zaphiris, P., & Wilson, S. (2010). Computer Games and Sociocultural Play: An Activity Theoretical Perspective. *Games and Culture*, *5*, 335–353.

Balicer, R. D. (2007). Modeling Infectious Diseases Dissemination Through Online Role-Playing Games. *Epidemiology*, *18*(2), 260–261. http://doi.org/10.1097/01.ede.0000254692.80550.60

Bardzell, S., Bardzell, J., Pace, T., & Reed, K. (2008). Blissfully productive: grouping and cooperation in World of Warcraft instance runs. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (pp. 357–360). New York, NY, USA: ACM Press. http://doi.org/10.1145/1460563.1460621

Bartle, R. (1996). Hearts, clubs, diamonds, spades: players who suit MUDs. *Journal of MUD Research*. Retrieved from http://www.mud.co.uk/richard/hcds.htm

Benesch, S. (2014). Defining and diminishing hate speech, State of the World's Minorities and Indigenous Peoples 2014, pp. 14-25.

Bergstrom, K. (2011, October). *Hulkageddon: The polarization of play in EVE Online*. Presented at the Association of Internet Researchers, Seattle, WA.

Bergstrom, K. (2016). Imagined Capsuleers: Reframing Discussions about Gender and EVE Online. In M. Carter, K. Bergstrom, & D. Woodford, *Internet Spaceships are Serious Business: An EVE Online Reader* (pp. 148–163). University of Minnesota Press.

Boman, M., & Johansson, J. S. (2007). Modeling Epidemic Spread in Synthetic Populations — Virtual Plagues in Massively Multiplayer Online Games. In *Situated Play, Proceedings of DiGRA 2007 Conference* (pp. 357–361). Tokyo, Japan.

Braithwaite, A. (2013). 'Seriously, get out': Feminists on the forums and the War(craft) on women. New Media & Society 0(0): 1-16. First published online June 12, 2013: 10.1177/1461444813489503.

Calvert, C. (1997). Hate Speech and Its Harms: A Communication Theory Perspective. Journal of Communication 47(1): 4-19. First published online February 7, 2006: 10.1111/j.1460-2466.1997.tb02690.x.

Chen, M. (2008). Communication, Coordination, and Camaraderie in World of Warcraft. *Games and Culture*, *4*(1), 47–73. http://doi.org/10.1177/1555412008325478

Chen, M. (2012). *Leet Noobs: The Life and Death of an Expert Player Group in World of Warcraft*. New York: Peter Lang.

Chesney, T., Coyne, I., Logan, B., & Madden, N. (2009). Griefing in virtual worlds: causes, casualties and coping strategies. *Information Systems Journal*, *19*(6), 525–548. http://doi.org/10.1111/j.1365-2575.2009.00330.x

de Castell, S., Jenson, J., Taylor, N., & Thumlert, K. (2014). Re-thinking foundations: Theoretical and methodological challenges (and opportunities) in virtual worlds research. *Journal of Gaming & Virtual Worlds*, *6*(1), 3–20. http://doi.org/10.1386/jgvw.6.1.3_1

de Castell, S., Taylor, N., Jenson, J., & Weiler, M. (2012). Theoretical and methodological challenges (and opportunities) in virtual worlds research (p. 134). ACM Press. http://doi.org/10.1145/2282338.2282366

Decker, S. A., & Gay, J. N. (2011). Cognitive-bias toward gaming-related words and disinhibition in World of Warcraft gamers. *Computers in Human Behavior*, *27*(2), 798–810. http://doi.org/10.1016/j.chb.2010.11.005

Ducheneaut, N., Yee, N., Nickell, E., & Moore, R. J. (2006). "Alone together?": exploring the social dynamics of massively multiplayer online games (p. 407). ACM Press. http://doi.org/10.1145/1124772.1124834

Eklund, L., & Johansson, M. (2010). Social Play? A study of social interaction in temporary group formation (PUG) in World of Warcraft. In *Proceedings of DiGRA Nordic 2010: Experiencing Games: Games, Play, and Players*. Retrieved from http://www.digra.org/dl/display_html?chid=10343.55072.pdf

Eklund, L., & Johansson, M. (2013). Played and Designed Sociality in a Massive Multiplayer Online Game. *ELUDAMOS Journal for Computer Game Culture*, *7*(1), 35–54.

Ghanea, N. (2013). Intersectionality and the Spectrum of Racist Hate Speech: Proposals to the UN Committee on the Elimination of Racial Discrimination. Human Rights Quarterly 35(4): 935-954. First published November 2013: 10.1353/hrq.2013.0053.

Goffman, E. (1990). *The presentation of self in everyday life*. New York [N.Y.]: Doubleday.

Gray, K. (2012a). Deviant Bodies, Stigmatized Identities, and Racist Acts: Examining the Experiences of African-American Gamers in Xbox Live. New Review of Hypermedia and Multimedia 18(4): 261-276. First published online: December 3, 2012: 10.1080/13614568.2012.746740.

Gray, K. (2012b). Intersecting Oppressions and Online Communities: Examining the experiences of women of color in Xbox Live. Information, Communication & Society 15(3): 411-428. First published online: December 19, 2011. 10.1080/1369118X.2011.642401.

Gray, K. (2012c). Collective Organizing, Individual Resistance, or Asshole Griefers? An Ethnographic Analysis of Women of Color in Xbox Live. Ada: A Journal of Gender, New Media, and Technology 2. First published online: June 2013. 10.7264/N3KK98PS.

Kou, Y., & Nardi, B. (2013). Regulating anti-social behavior on the Internet: The example of League of Legends. iConference 2013 Proceedings, Fort Worth, TX: February 12-15, 2013, pp. 616-622. 10.9776/13289.

Lewis, C., & Wardrip-Fruin, N. (2010). Mining game statistics from web services: A World of Warcraft Armory Case Study (pp. 100–107). ACM Press. http://doi.org/10.1145/1822348.1822362

Li, D. D., Liau, A. K., Gentile, D. A., Khoo, A., & Cheong, W. D. (2012). Construct and Predictive Validity of a Brief MMO Player Motivation Scale: Cross-sectional and longitudinal evidence based on Singaporean young gamers. *Journal of Children and Media*, 1–20. http://doi.org/10.1080/17482798.2012.712918

Lin, H., & Sun, C.-T. (2005). The "White-eyed" Player Culture: Grief Play and Construction of Deviance in MMORPGs. In *Proceedings of DiGRA 2005 Conference: Changing Views – Worlds in Play*. Retrieved from http://summit.sfu.ca/system/files/iritems1/238/5922543c8cba0a282491dbfdfb17.doc

Liu, M., & Peng, W. (2009). Cognitive and psychological predictors of the negative outcomes associated with playing MMOGs (massively multiplayer online games). *Computers in Human Behavior*, *25*(6), 1306–1311. http://doi.org/10.1016/j.chb.2009.06.002

Lofgren, E. T., & Fefferman, N. H. (2007). The untapped potential of virtual game worlds to shed light on real world epidemics. *The Lancet Infectious Diseases*, *7*(9), 625–629. http://doi.org/10.1016/S1473-3099(07)70212-8

Meyers, D. (2007). Self and selfishness in online social play. In *Situated Play, Proceedings of DiGRA 2007 Conference*.

Nardi, B., & Harris, J. (2006). Strangers and friends: collaborative play in World of Warcraft. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* (pp. 149–158). ACM Press. http://doi.org/10.1145/1180875.1180898

Nardi, B., Ly, S., & Harris, J. (2007). Learning Conversations in World of Warcraft (pp. 79–79). IEEE. http://doi.org/10.1109/HICSS.2007.321

O'Sullivan, P. B., and Flanagin, A. J. (2003). Reconceptualizing 'flaming' and other problematic messages. New Media & Society 5(1): 69–94.

Paul, C. A. (2010). Welfare Epics? The Rhetoric of Rewards in World of Warcraft. *Games and Culture*, *5*(2), 158–176. http://doi.org/10.1177/1555412009354729

Pulos, A. (2013). Confronting Heteronormativity in Online Games: A Critical Discourse Analysis of LGBTQ Sexuality in World of Warcraft. Games and Culture 8(2): 77-97.

Ratan, R. A., Chung, J. E., Shen, C., Williams, D., & Poole, M. S. (2010). Schmoozing and Smiting: Trust, Social Institutions, and Communication Patterns in an MMOG. *Journal of Computer-Mediated Communication*, *16*(1), 93–114. http://doi.org/10.1111/j.1083-6101.2010.01534.x

Salter, A. and Blodgett, B. (2012). Hypermasculinity & Dickwolves: The Contentious Role of Women in the New Gaming Public. Journal of Broadcasting & Electronic Media 56(3): 401-416.

Shen, C. (2010). *The patterns, effects and evolution of player social networks in online gaming communities* (Ph.D. Dissertation). University of Southern California.

Sherr, I. (2014). Player Tally for League of Legends surges. The Wall Street Journal. First published January 27, 2014. http://blogs.wsj.com/digits/2014/01/27/player-tally-for-league-of-legends-surges/.

Silverman, M., & Simon, B. (2009). Discipline and Dragon Kill Points in the Online Power Game. *Games and Culture*, *4*(4), 353–378. http://doi.org/10.1177/1555412009343572

Sinders, C. (2015). I Was On One Of Those Canceled SXSW Panels. Here Is What Went Down. Slate. First published October 29, 2015. http://www.slate.com/articles/double_x/doublex/2015/10/sxsw_canceled_panels_here_is_what_happened.html.

Stetina, B. U., Kothgassner, O. D., Lehenbauer, M., & Kryspin-Exner, I. (2011). Beyond the fascination of online-games: Probing addictive behavior and depression in the world of online-gaming. *Computers in Human Behavior*, *27*(1), 473–479. http://doi.org/10.1016/j.chb.2010.09.015

Suznjevic, M., & Matijasevic, M. (2010). Why MMORPG players do what they do: relating motivations to action categories. *International Journal of Advanced Media and Communication*, *4*(4), 405. http://doi.org/10.1504/IJAMC.2010.036838

The British Institute of Human Rights Council. (2012). Mapping Study on Projects against Hate Speech Online. Young People Combating Hate Speech Online. http://www.coe.int/t/dg4/youth/Source/Training/Training_courses/2012_Mapping_projects_against_Hate_Speech.pdf (accessed 29 August 2014).

United Nations General Assembly. International Covenant on the Elimination of All Forms of Racial Discrimination. United Nations. http://www.ohchr.org/EN/ProfessionalInterest/Pages/CERD.aspx (accessed 04 August 2014).

Valkyrie, Z. C. (2011). Cybersexuality in MMORPGs: Virtual Sexual Revolution Untapped. *Men and Masculinities*, *14*(1), 76–96.

Whitty, M. T., Young, G., & Goodings, L. (2011). What I won't do in pixels: Examining the limits of taboo violation in MMORPGs. *Computers in Human Behavior*, *27*(1), 268–275. http://doi.org/10.1016/j.chb.2010.08.004

Williams, D. (2010). The Mapping Principle, and a Research Framework for Virtual Worlds. *Communication Theory*, *20*(4), 451–470. http://doi.org/10.1111/j.1468-2885.2010.01371.x

Williams, D., Ducheneaut, N., Xiong, L., Zhang, Y., Yee, N., & Nickell, E. (2006). From Tree House to Barracks: The Social Life of Guilds in World of Warcraft. *Games and Culture*, *1*(4), 338–361. http://doi.org/10.1177/1555412006292616

Yee, N. (2006). Motivations for Play in Online Games. *CyberPsychology & Behavior*, *9*(6), 772–775. http://doi.org/10.1089/cpb.2006.9.772

Yee, N. (2008). Befriending Ogres and Wood-Elves: Relationship Formation and The Social Architecture of Norrath. *Game Studies*, *8*(2), [Online].

Yee, N., Ducheneaut, N., Yao, M., & Nelson, L. (2011). Do men heal more when in drag? In *Proceedings of the 2011 annual conference on Human factors in computing systems (CHI '11).* (pp. 773–776). ACM Press. http://doi.org/10.1145/1978942.1979054

---

[i] https://www.datasociety.net/pubs/oh/Online_Harassment_2016.pdf
[ii] https://en.wikipedia.org/wiki/Wikipedia:Wikipedians#Number_of_editors
[iii] https://blog.wikimedia.org/2017/02/07/scaling-understanding-of-harassment/
[iv] https://www.splcenter.org/fighting-hate/extremist-files/individual/david-horowitz
[v] https://russell.house.gov/uploadedfiles/waste_watch_no_3_oct_20_2015.pdf